Bibliometric Study in Text Mining and Maintenance

Virgínia Siqueira Gonçalves[1], Elias Rocha Gonçalves Júnior[2], Álvaro de Azeredo Araújo de Carvalho[3]

[1, 2, 3]Universidade Candido Mendes, Centro de Pesquisa Candido Mendes,
Rua Anita Peçanha, 100, Campos dos Goytacazes, Rio de Janeiro, Brazil

Abstract: *This research aims to present a bibliometric study which goal is to identify the works that deal with the maintenance and using of data mining techniques. To the current study, articles previously published in the SCOPUS base were used as data sources. This article presents and justifies the using of text mining in the context of maintenance and how it can be applied in the decision-making. It was found 114 works of many parts of the world. Thus, it can be concluded that this theme is global, current and relevant. The study allowed us to identify the main used methods.*

Keywords: Text mining, Bibliometrics, Maintenance.

1. Introduction

A growing number of non structured data storage applications have grown widely due to the progress of information technologies, once these applications have valuable information, either behavior patterns, trends or abnormalities which are used in decision-making [1].

Just a small fraction of the analyzed files are considered relevant for a determinate purpose. But, when the content of each text is not known, the action of extracting any useful informations becomes almost impossible. Thus, methods and techniques were used to analyze files and sort them according to labels established previously [2].

The task identifying and keeping up with important aspects for decision-making, against the huge amount of maintenance orders is difficult. Mainly because of the difficulty to manipulate them while they do not have a structured data format, which, in most cases, are found in a semi-structured format. Once these data have grown exponentially, and are available in many platforms, it is necessary to use recovery techniques and treatment, which goal is to analyze them in a consistent way

These techniques are called of Data Mining, which derives of Data Mining techniques, once both aim to extract useful information in non-structured data or semi-structured data and these data are not easily manipulated [3].

In Data Mining there are few techniques, which are: Machine Learning, Computational Intelligence, Information Recovery, Cognitive Science, Information Extraction, Natural Language Processing, among others [4].

It is known that every Corporation in the world has a huge amount of text files, as it is the most natural way of storing information. A part of these data is stored or available in the internet, creating big data collections, like: machine specification, errors reports, fault diagnostics, maintenance orders, abstracts, notes, electronic correspondence report cards, purchase orders and every type of publication [2].

Increasing productivity and eliminating costs are terms that are in tremendous force in today's businesses, so in competitive organizations, maintenance has a strategic function. [5] says that time mining can also be viewed as a competitive advantage to organizations, as well as support for decision making, with indicative of successes and failures, since it allows to extract useful knowledge through raw unstructured data.

[5] proposes a model in which the references are arranged in a table in order to provide a selection of academic papers to a new researcher in an area of knowledge. This process is divided into six steps: Define the research sample; Search with keywords in the sample; Identify the journals that publish the most in the area; Identify the authors with the greatest publication; Raise production chronology; and Identify the "starting nucleus" for a bibliographic search.

In this context, the present bibliometric study aims to identify studies that address data mining related to the maintenance theme, aiming at analyzing and identifying the authors, periodicals, countries, type of publication and areas of knowledge with the largest publications. It was tried to present ten works of practical application of the studied subject.

2. Methodology

A bibliometric study was carried out in order to identify the works that deal with maintenance using data mining techniques, as well as to present a vision of the scientific production. The database consulted in the preparation of this work was SCOPUS, since it contains not only academic articles but other types of publications. Exclusion filters were not used to make the document range of the area found larger. This study was based on the 10 main data provided by the database, ie for each category the first ten data provided by the database was considered.

Only the title, keyword, and summary fields were included as the search field. This study covered all areas of knowledge in the database. The "AND" operator ensures that the publications found contain all the search words, while the use of the "OR" operator enables one or the other term.

The research used the set of keywords or key terms ("Text Mining" AND "Maintenance"), reaching 114 results. This

will be illustrated by graphs of publications by year, by country, area of knowledge, the authors who publish the most and the journals with the greatest number of publications.

Figure 1 shows the number of publications per year. From 1998 to 2005, the periodicity of publications on this subject was low, so it will not be represented in the following figure.
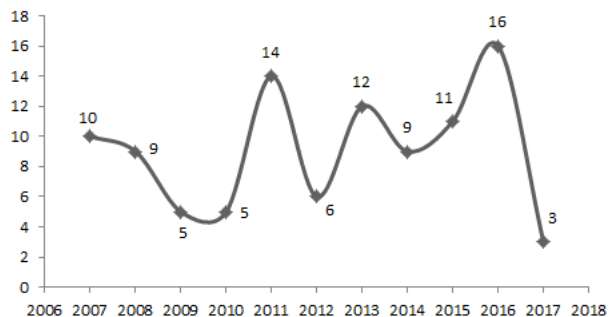

Figure 1: Publications identified at the Scopus Base.

It can be observed that the years 2011, 2013 and 1016 were the ones with the greatest volume of publication, 2 in 2017, until the current date of the survey, there was a considerable drop in relation to the other years analyzed.

In terms of journals, what stands out most is the Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics with 13 publications in the field, a number well away from the other journals, followed by Autotestcon Proceedings, how can be observed in Table 1. As previously mentioned, only the 10 main journals were selected for analysis.

Table 1: Periodicals with the highest number of publications of the records found

Affiliation	Documents
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	13
Autotestcon Proceedings	3
ACM International Conference Proceeding Series	2
Ceur Workshop Proceedings	2
Empirical Software Engineering	2
IEEE International Conference On Software Maintenance Icsm	2
Information Sciences	2
Plos One	2
Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining	2
Advanced Engineering Informatics	1

Analyzing the publications by author, Figure 2, we note that two stand out for having 6 publications each, these being Witte, R. and Rilling, J.
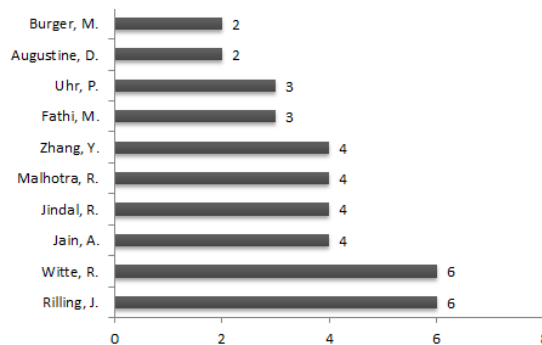

Figure 2: Distribution of the authors of the records found.

In Table 2 it is possible to observe the affiliations / institutions that are most interested in the subject in question, it is noted that University of Concordia has the largest number of publications in the area with 6, followed by Delhi Technological University with 4 publications.

Table 2: Affiliations with the highest number of publications of the records found

Affiliation	Documents
University of Concordia	6
Delhi Technological University	4
Karlsruhe Institute of Technology	4
Naval Air Systems Command	3
Zhejiang University	3
Beijing Jiaotong Daxue	3
Universitat Siegen	3
University of Illinois at Urbana-Champaign	2
Politechnika Warszawska	2
University of Manchester	2

Figure 3 illustrates that the United States is the country that most researches on the subject, counting 30 publications against 12 of China and India and 11 of Germany. It can be observed that there is no presence of Brazil in the relation of the main countries with related publications.
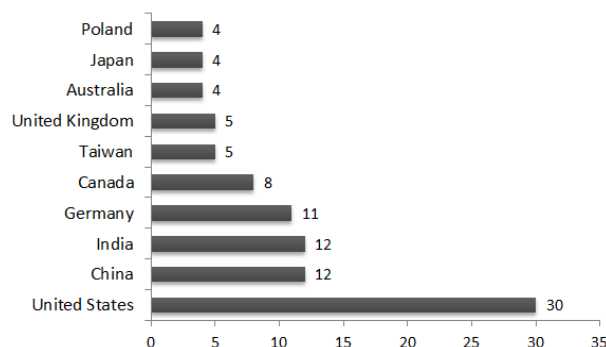

Figure 3: Countries with the highest number of publications of the registry found.

The articles analyzed were grouped in 10 areas of knowledge, by the SCOPUS database itself, as shown in Figure 4: Computer Science; Engineering; Mathematics; Decision Science; Medicine; Biochemistry, Genetics and Molecular Biology; Business, Management and Accounting; Energy; Agriculture and Biological Sciences.

The area of Computer Science stands out with 83 publications in the area, more than half of the second area of
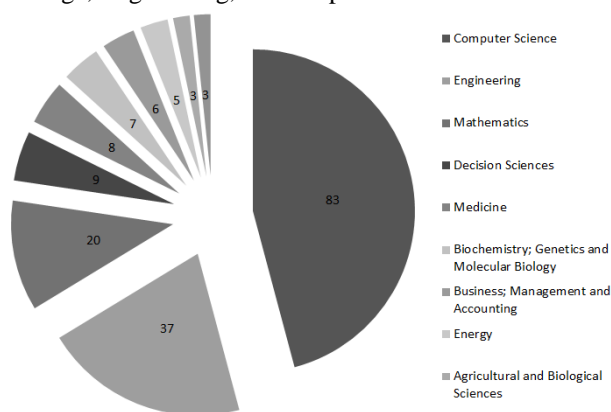
knowledge, Engineering, with 37 publications.

Figure 4: Knowledge areas with the largest number of publications in the registry found

When it comes to the means of publication, the SCOPUS database grouped into types, as shown in Figure 5. It is noted that congress papers present more than the sum of the other types of publication.
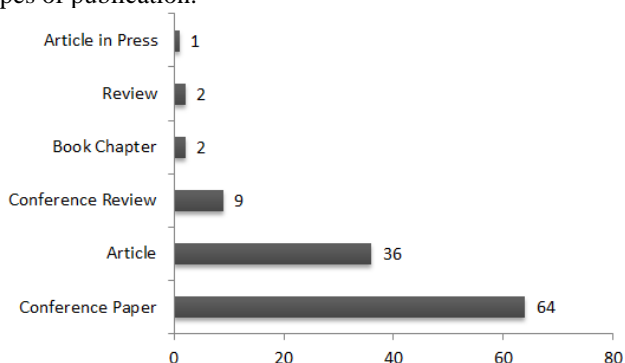

Figure 5: Types of publications of the registry found

3. Results and Discussions

For this stage, the 114 documents were analyzed and 10 articles were selected among the others, based on the analysis of their scientific summaries. These, arranged to follow in Table 3, are those considered by the author as those that best correlate text mining techniques and the theme maintenance.

Table 3: Selected articles

Title	Author	Year	Citations
Predicting Software Maintenance Effort by Mining Software Project Reports Using Inter-Version Validation	Jindal, R., Malhotra, R., Jain, A.	2016	0
Predicting Software Maintenance effort using neural networks	Jindal, R., Malhotra, R., Jain, A.	2015	0
Extracting failure time data from industrial maintenance records using text mining	Arif-Uz-Zaman, K., Cholette, M.E., Ma, L., Karim, A.	2017	0
The management of the knowledge evolution by using text mining techniques	Nora, T., Mokhtar, S., Simonet, M.	2009	1
Bi-level Feature Extraction-Based Text Mining for Fault Diagnosis	Wang, F., Xu, T.,	2017	3

Title	Author	Year	Citations
of Railway Systems	Tang, T., Zhou, M., Wang, H.		
Text mining based fault diagnosis of vehicle on-board equipment for high speed railway	Zhao, Y., Xu, T.-H., Wang, H.-F.	2014	5
Knowledge Modeling in Construction of Technical Management System for Large Warehousing Facilities	Gajzler, M.	2015	1
Physics of failure, predictive modeling & data analytics for LOCA frequency	O'Shea, N., Pence, J., Mohaghegh, Z., Kee, E.	2015	1
Verification of text mining techniques accuracy when dealing with urban buses maintenance data	Marzec, M., Uhl, T., Michalak, D.	2014	0
Automotive diagnosis typo correction using domain knowledge and machine learning	Huang, Y., Murphey, Y.L., Ge, Y.	2013	3

Already selected articles, a brief summary of each will be presented.

Software changes are inevitable when it comes to a dynamic and active environment, where user expectations and requirements tend to change rapidly, so software needs to be constantly updated. This update is performed in terms of the number of rows in the code (LOC) that may have been modified, inserted, and / or deleted from an earlier version to the next. These changes must be maintained in reports that constitute the identification and description of the defect, presenting the factor that may have caused the defect occurred in the previous version, demonstrating the reason for the change of the LOC, either by presenting the reason for which it was inserted, deleted or modified. In the maintenance phase of the software, much effort is made to correct the errors identified. Thus, [6] proposed to predict maintenance efforts by analyzing defect reports using text mining techniques and then developing a prediction model using appropriate machine learning algorithms, Multi-Layer Perceptron (MLP), Radial-Basis Function (RBF) network and Decision Tree (DT). For this work, we considered the changes in three successive versions of the "MMS" application package of the Android operating system and performed the validation between the versions, that is, when the 'v' version is validated, it becomes' v + 1 '. The model performance was evaluated using the receiver operational characteristics (ROC) analysis. The results indicated that the predicted model in the 'MMS' version 4.0 using the MLP algorithm showed good results when validated in the 'MMS' version 4.1. On the other hand, the performance of the RBF and DT algorithms has been consistently average in predicting the maintenance effort.

Maintenance is an important stage in the software development life cycle, which begins after delivery to the customer. This is a laborious activity, since it is necessary to make changes in a running software. One of the main problems was to predict the cost and effort associated with maintenance activities, needing to be analyzed for effective resource allocation and decision making. For this, [7] developed a model based on text mining techniques using

machine learning method named Radial Base Function of the neural network. Text mining techniques were applied to identify the relevant attributes of maintenance reports and their relationship to the prediction of maintenance efforts. The proposed model is validated using the application package 'Browser' of the Android operating system. The Receiver Operational Characteristics (ROC) analysis is performed to interpret the results obtained from the prediction model using the Area Under Curve (AUC) value, sensitivity, and an appropriate threshold criterion known as the cutoff point. From the results, it was noticed that the performance of the model is dependent on the number of words considered for classification and, therefore, shows the best results in relation to the 100 main words. Performance is independent of effort category type.

[8] believes that reliability modeling requires a precise failure time, but when one speaks of the actual industrial cases, this data is placed in banks that do not allow it. Two sets of data are commonly available: work orders (OS), which detail maintenance activities, and downtime data (DD), which details when the feature has been disconnected. Both are incomplete, since we want to know if the time each resource is inactive was due to failure or previously planned activity. Thus, the authors propose the use of text mining to extract accurate failure time data from the OS's and DD's. From the descriptions of the texts of the OSs a dictionary of keywords was created applied to each DD event being classified as 'failed' or 'not failed'. The proposed method identifies the downtime events whose descriptions are consistent with unplanned urgent requests. This has been applied to the maintenance data sets of Australian electricity and sugar processing companies. The analysis of the text of the identified failure events seems to confirm the accurate identification of the failures in DD. The authors expect the results to be immediately useful in improving the estimation of failure times (and hence reliability models) for real-world assets.

Maintaining the domain ontology or a knowledge model is an essential step after the onset of change. Several studies propose methodologies to perform the maintenance of an ontology, but few are for ontologies created from texts. Knowing that text mining techniques provide good results when text processing is done for the purpose of modeling or classification, [9] proposed the use of mining techniques to perform the maintenance of an ontology that represents a dynamic domain, according to an analysis of textual changes and their effects on the corresponding ontology, using two algorithms: CAH Algotithm and Apriori. In order to perform the ontology maintenance, two types of coherence have been defined: Internal coherence means the ontological coherence of the entities, and External coherence means to maintain the dependence of the ontology in question with other ontologies (specific or generic). As a conclusion, he obtained that the study of the changes that can intervene in the texts, as well as the formalism of the representation that was adopted, facilitates the maintenance of the ontology.

According to [10], due to the accumulation of data regarding the maintenance of operation in an on-board vehicle equipment, it plays an important role in the diagnosis and prognosis of failures. However, the natural language used in the maintenance data makes the diagnosis of failures a challenge, since they have semantics and irregular characteristics. It shows that some researchers have proposed to introduce text mining in order to solve this problem, however, they did not obtain good results because they did not have previous knowledge. This work aimed to present the fault diagnosis based on the Latent Dirichlet Allocation (LDA) method and the vector machine support. First, the frequency and frequency method of the inverse topic is proposed to extract prior knowledge about the fault symptom, which is then integrated into the basic LDA method. Then, classifier resources are extracted, from the LDA model, from the maintenance records. Third, a hierarchical classification model based on the vector machine support and the resource fusion method that is used to diagnose failures has been proposed. Finally, the frequency method is introduced to evaluate the performance of the proposed model with real data of the high-speed rail system at the Guangzhou Railway Corporation. As a result, they observed that the experiments showed that the proposed method surpasses the method of text mining performed without previous knowledge and other common methods of fault diagnosis.

Due to the irregularity and arbitrariness used in tracking failure records, [11] proposed a method of fault diagnosis of a vehicle equipment on board the high speed rail signal system (VOBE) for fault analysis and resource extraction, based on the Bayesian network. For the adequacy of the Bayesian network, characteristics and domain knowledge, a new algorithm was proposed called HDBN_SL (Bayesian Hierarchical Diagnostic Networks - Learning Structure) applicable to VOBE. The experimental analysis was performed based on field data from Wuhan-Guangzhou high-speed rail signaling systems. The test results showed that the mining of text mining features and the failure diagnostic method provided better diagnostic accuracy.

[12] addresses the problem encountered in the technical management of large storage facilities from the aspect of building a decision support system or, more precisely, acquiring the necessary knowledge as a resource based on automated methods. This presents standard procedures that are in place in Poland regarding the technical maintenance of the building structures and the tools in the form of dedicated systems. It proposes the realization of a decision support system for the management of large storage facilities, creating a practical opportunity to simplify and reduce the work of the facilities manager. The author believes that there are limitations to text mining, since quantitative descriptions were avoided since they did not make sense out of context and instead qualitative descriptions were used. For the acquisition of knowledge, the system HASIFR (Hybrid Consultation System for industrial coating repairs) was created to classify materials and technological solutions used in different repair systems as useful or useless to repair damaged floor in specific circumstances. In maintenance techniques, it is a "first contact" device used when it is necessary to repair damaged floor, serving as a "virtual advisor". It concluded that a system focused on supporting the technical management of storage facilities, mixed /

hybrid approach to knowledge acquisition should be used, however, using only one knowledge source does not guarantee the reach of a pseudo-complete knowledge needed to manage a given installation, since each method has its limitation.

[13] presented a methodology of theoretical data as part of a research that is underway to integrate physical fault theory (PoF) and data analysis to be applied in the probabilistic risk assessment (PRA) of complex systems, besides the status of application of the methodology proposed for the location-specific refrigerant loss accident estimation (LOCA), which is a critical initiation event in PRA and one of the challenges of risk resolution for generic safety problems (GSI-191). The adopted methodology uses a predictive causal modeling together with the sensitivity and uncertainty analysis to find the most important contributing factors in the PoF models of failure mechanisms. When a methodology is based on a model that uses classification techniques of importance, the number of factors is reduced scientifically, focusing on a strategy of minimal quantification directed to critical factors instead of carrying out experiments with high costs and simulations that demand time for a range of factors, thus adding practicality and validity to the proposed methodology. This theoretical data method, with the evolution of the power of information sharing and computational power, when applied to a PRA, expands the classical data extraction and implementation approach for risk analysis. It used the data mining and text mining techniques in the data (academic literature, service data, laboratory and regulatory reports, expert opinion, maintenance logs, news and others) in order to extract risk and reliability information, in addition to of executing them in networks based on PoF theory. The theoretical data approach, based on the theory of PoF, avoids generating potentially erroneous results when using a purely numerical data approach, it is necessary to complete them with the physical textual factors and the precision of their causal relations. Quantified important and causal factors, verified and validated the causal model of PoF.

[14] believes that with the constant increase of maintenance costs, it is necessary to develop a good planning of these activities. Reviewing maintenance history records minimizes maintenance costs and prevents the downtime of a bus fleet. Since these records are in free text, their analysis requires an individual examination of their content, however, this process can be automated with the use of text mining techniques, knowing that the accuracy of the analysis will depend on the quality of the data and the methods which will be applied, therefore, you must test in advance and analyze your results. Text mining is especially important to aid in service decision making as it influences maintenance and safety costs. Thus, the authors aimed to determine if the existing and currently used mining methods are sufficiently accurate to be used in the classification of unstructured urban bus maintenance and repair data. As a conclusion it has obtained that the text mining presents excellent classification skills proposed, since this model has 99% accuracy and can be applied to support maintenance decisions.

[15] believes that the description of diagnoses after vehicle repair play a crucial role in cause and root analysis and in vehicle maintenance itself. However, because they are unstructured texts, do not have grammar, have spelling mistakes and auto-invented terminologies, it is a challenge to automatically retrieve and categorize this information. Thus, the present article sought to apply text mining in the diagnostics of vehicles, proposing and implementing an automatic system of detection and correction of typing, as well as a neural network classification to select good candidates to correct these errors. As a result, they have obtained that the system created by them outperforms state-of-the-art spell check systems.

4. Conclusion

The present work aimed to present a study on the application of the method and techniques of text mining applied to maintenance, indicating a collection of academic articles, considered by the author, important referring to the area of study.

It can be said that the model of bibliometry proposed by Costa (2010) is valid when directing a researcher on a certain area, since academic papers are gathered on a certain theme based on their relevance and credibility, in addition to presenting the authors who publish the most , journals that are more interested in the subject, countries that carry out research on the subject, related areas of knowledge, types of publications and institutions with the highest number of publications.

It was possible to observe during the research that there are no academic papers in Portuguese referring to the studied subject, all these were in the English language, being this one the most used for publication of academic works currently. Another reason for this is that the country with the largest number of publications is the United States (see Figure 3).

Based on the results and discussion, it is noted that text mining is a method that can be applied in several areas and applications as presented in this article, as an aid in decision making, extracting information, correcting misspellings, reducing costs, extraction of characteristics, indicate events occurred, maintenance of knowledge, among others.

5. Acknowledgements

References

[1] M. W. Berry, J. Kogan, "Text mining: applications and theory", [S.l.]: John Wiley & Sons, 2010.
[2] J. Han, M. Kamber, J. Pei, "Data mining: concepts and techniques: concepts and techniques", [S.l.]: Elsevier, 2011.

[3] R. Feldman, J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data", [S.l.]: Cambridge University Press, 2007.

[4] G. G. Chowdhury, "Natural language processing", Annual review of information science and technology, Wiley Online Library, v. 37, n. 1, p. 51–89, 2003.

[5] H. Costa, "Modelo para webibliomining: proposta e caso de aplicação", Rev. FAE, Curitiba, v13, n.1, p. 115-126, jan./jun.2010.

[6] R. Jindal, R. Malhotra, A. Jain, "Predicting Software Maintenance Effort by Mining Software Project Reports Using Inter-Version Validation", International Journal of Reliability, Quality and Safety Engineering, Vol. 23, No. 6, 2016.

[7] R. Jindal, R. Malhotra, A. Jain, "Predicting Software Maintenance effort using neural networks", 4th International Conference on Reliability, Infocom Technologies and Optimization: Trends and Future Directions, ICRITO 2015, 2015.

[8] K. Arif-uz-zaman, M. E. Cholette, L. Ma, A. Karim, "Extracting failure time data from industrial maintenance records using text mining", Advanced Engineering Informatics, 33, pp. 388-396, 2017.

[9] T. Nora, S. Mokhtar, M. Simonet, "The management of the knowledge evolution by using text mining techniques", 9th International Conference on Knowledge Management and Knowledge Technologies and Proceedings of I-SEMANTICS 2009 - 5th International Conference on Semantic Systems pp. 663-668, 2009.

[10] F. Wang, T. Xu, T. Tang, M. Zhou, H. Wang, "Bi-level Feature Extraction-Based Text Mining for Fault Diagnosis of Railway Systems", IEEE Transactions on Intelligent Transportation Systems, 18(1), 7453147, pp. 49-58, 2017.

[11] Y. Zhao, T. H. Xu, H. F. Wang, "Text mining based fault diagnosis of vehicle on-board equipment for high speed railway", 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 6957803, pp. 900-905, 2014.

[12] M. Gajzler, "Knowledge Modeling in Construction of Technical Management System for Large Warehousing Facilities", Procedia Engineering, 122, pp. 181-190 , 2015.

[13] N. O'Shea, J. Pence, Z. Mohaghegh, E. Kee, "Physics of failure, predictive modeling & data analytics for LOCA frequency", Proceedings - Annual Reliability and Maintainability Symposium 2015-May,7105125, 2015.

[14] M. Marzec, T. Uhl, D. Michalak, "Verification of text mining techniques accuracy when dealing with urban buses maintenance data, Diagnostyka", 15 (3), pp. 51-57, 2014.

[15] Y. Huang, Y. L. Murphey, Y. Ge, "Automotive diagnosis typo correction using domain knowledge and machine learning", Proceedings of the 2013 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2013 - 2013 IEEE Symposium Series on Computational Intelligence, SSCI 2013 6597246, pp. 267-274, 2013.

Author Profile

Virgínia Siqueira Gonçalves Master student in Production Engineering from Universidade Candido Mendes (UCAM). Master student in Materials Engineering and Science from the Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF). Graduated in Mechanical Engineering from Universidade Candido Mendes (UCAM). Shee was a member of the Organizing Committee of the XXII National Congress of Mechanical Engineering Studies (CREEM). As a professional, she was a lecturer and lecturer at the PBF - Pink and Blue Freedom School (Campos dos Goytacazes / RJ). She interned at Petróleo Brasileiro S & A - PETROBRAS in the area of Integrity Maintenance (Macaé/RJ). Flexible Tubes Design Engineering at National Oilwell Varco - NOV Flexibles (Super Porto do Açu/RJ). She is currently a lecturer at ITECAM - Campos Technological Institute, based at Universidade Candido Mendes - Campos (Campos dos Goytacazes/RJ) and at IFF – Instituto Federal Fluminense Campus Campos dos Goytacazes.

Elias Rocha Gonçalves Júnior Master student in Production Engineering from Universidade Candido Mendes (UCAM). Master student in Materials Engineering and Science from the Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF). Postgraduate in Teaching of Higher Education from the Faculdade Integrada de Araguatins (FAIARA). Graduated in Mechanical Engineering from Universidade Candido Mendes (UCAM). Researcher at the Inter-Institutional Research Group for Municipal / Regional Development - ITEP / UENF / UNIFLU, Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF) in partnership with the Centro Universitário Fluminense (UNIFLU). He is currently Professor Instructor I at Universidade Candido Mendes - Campos (UCAM). He has experience and research in Mechanical Engineering and Production Engineering, with emphasis on Renewable Energies, Industrial Maintenance, Data Mining, Strategic Planning, Combustion Engines, Special Education and Sustainable Management. He is lecturer, author of books, occupant of the 32nd chair of the Pedralva Letras and Arts Academy, and member of the Brazilian Association of Engineering and Mechanical Sciences (ABCM).

Álvaro de Azeredo Araújo de Carvalho. Graduated in Mechanical Engineering from Candido Mendes University - Campos. He was a member of the BAJA SAE Team at Candido Mendes - Campos University. As a professional, he was Technical Design Monitor I, II and III at the Candido Mendes - Campos University and trainee at Technip in the area of Methods and Process Industrialization (Super Port of Açu / RJ). He currently holds an MBA in Project and Business Management at Candido Mendes - Campos University, an English course at Number 1 (Campos dos Goytazas / RJ) and a French course at Candido Mendes - Campos University.